16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

PROCESSING OF LOW QUALITY DOCUMENT IMAGES: ISSUES AND DIRECTIONS.

M. Cheriet and R. Farrahi Moghaddam

Synchromedia Laboratory for Multimedia Communication in Telepresence,
École de Technologie Supérieure, Montréal, (QC) H3C 1K3 Canada,
mohamed.cheriet@etsmtl.ca, reza.farrahi-moghaddam.1@ens.etsmtl.ca)

ABSTRACT

Issues facing document image analysis and recognition are discussed based on the quality and complexity of images. Special attention is paid to low-quality images of ancient manuscripts. Because of the complex content of this type of document, which usually contains several layers of information in the same scale levels, the definition of degradation must be reconsidered. This opens up new challenges for the modeling of document degradation. Also, discussed is the development of appropriate restoration methods for handling degradation. The advantages of preprocessing a document to remove some of the unwanted layers of information in the document image in order to improve its quality are considered, using currently available or new paradigms.

1. INTRODUCTION

"The heritage of the past is the seed that brings fourth the harvest of the future."

The heritage of a community defines the character and uniqueness of that community, and many of its customs and habits can be traced back to their origins through an analysis of the corresponding heritage in the arts, documents, etc. However, because of the high degree of degradation suffered by very old documents and artifacts, analyzing them is a difficult task.

Document imaging is a very powerful approach to sharing and distributing valuable and historic documents. However, in many cases, search and other text-related functions are essential for full utilization of archives. The content-based archiving of very old documents is the best way to achieve this. However, doing so manually is always a costly and time-consuming process. Also, it is subject to human error and bias. There are several automatic and powerful optical character and text recognition (OCR/ICR) algorithms available for document imaging, but, to be effective, they require fairly clean, readable inputs.

There is a very large number of degradation types in very old documents, bleed-through being one of the most challenging. The problem is very much harder to address when the same ink color has been used on both sides of a document, or in the case of gray-scale documents. There are many approaches to solving this sort of problem [1, 2, 3, 4]. For example, binarization methods are used to restore documents with bleed-through defects [1, 5, 6]. In more general approaches, statistical methods, such as Independent Component Analysis (ICA) and Blind Source Separation (BSS) [4, 7], are applied. Neural networks are also considered [8]. Moreover, there are other methods which combine of several techniques, such as segmentation and inpainting [9, 2, 10]. In this work, we compare the statistical methods, which are the most promising, with our novel approach based on the diffusion methods. Our comparison is conducted from a fundamental point of view, to enable a better understanding of the advantages and disadvantages of the methods. Also, in addition to providing real samples, which we obtained from [11, 2], a degradation model is developed which is capable of generating an unlimited number of document images degraded by bleed-through. This model is discussed in the next section. To our knowledge, there is only one other degradation model for this type of defect, and it is based on blurring and mixing techniques [12, 13]. Finally, possible directions for the restoration and enhancement of very old documents are offered which benefit from the advantages of various methods.

2. METHODS

In this section, we briefly introduce the methods used in the study.

- **Statistical Methods:** In BSS applications, the premier tool is the statistical approach, in which the input images are assumed to be signals obtained by mixing a number of sources. In this way, the input images are analyzed as one-dimensional arrays from the start, which means that the two dimensional correlation of the input images is ignored. The best approach is obviously a statistical method, such as Independent Component Analysis (ICA), especially when the sources are assumed to be independent. Although different types of functions are used in ICA methods, the basic idea is simple: there is a cost function which determines the degree of Independence of the computed sources. By maximizing this cost function, the best estimation for the sources is obtained. Also, these methods assume a linear relation between the sources and the inputs.
- **Diffusion Methods:** These methods are based on the existence of a spatial correlation between the data of neighboring pixels, and so each pixel is processed (restoration, enhancement, etc.) using the information of the pixels surrounding it. In other words, it is assumed that, by some degradation process (such as the application of blur, the addition of noise, etc.), the true image data is destroyed, and the data must be corrected via exchange of information between neighbors. A direct consequence of this is that all weak structures are subject to removal, which makes these methods very aggressive. Obviously, diffusion methods in this form are not applicable to source separation problems. However, for double-sided document images, we can modify them

to make them applicable to two-source separation problems, in which there are two input images that are actually mixtures of two unknown sources (the images of the recto and verso sides of a document). For visualization purposes, we imagine that the input images are placed on two plates. At the end of computations, the two sources will appear on the same plates. In addition to the usual diffusion processes on each plate, we add some diffusion processes from the other plate, but in the reverse direction, and call it a double-sided flow-based diffusion method (DFDM). In doing so, we arrive at a method in which these reverse diffusion processes cancel out the effects of the real physical degradation processes that occur over time. These additional diffusion processes actually separate the recto and verso side information by pushing the interference patterns to the background. For better suppression of the interference patterns and a more uniform background, another diffusion process is included which transfers data from an estimated background plate for each side of the document to the corresponding plate. This process not only results in a uniform and fluctuation-free background, but it also speeds up the removal of interference patterns by filling up the patterns as they are pushed to the background by the reverse diffusion processes. In other words, the reverse processes pass on the interference patterns to the background diffusion process. All the diffusion processes that participate in this method can be summarized in the following equation:

$$\frac{\partial u}{\partial t} = \mathrm{div}_f\left(c_f \nabla u\right) + \mathrm{div}_{r,f}\left(c_{r,f} \nabla_r u\right) + \mathrm{div}_{b,f}\left(c_{b,f} \nabla_b u\right) \tag{1}$$

where u represents the image data, and c represent the diffusion coefficients. The subscripts r and b stand for the reverse and background diffusion respectively. The flow field is denoted by f and represents a classifier from a global point of view, and it helps in discriminating between the boundary pixels and the inside pixels, as well as better preserving the edges and boundaries.

If we use normal diffusion processes between the recto and verso plates in the diffusion method, the model simulates the actual degradation of documents due to ink seepage and bleed-through. This degradation model is used in the following examples.

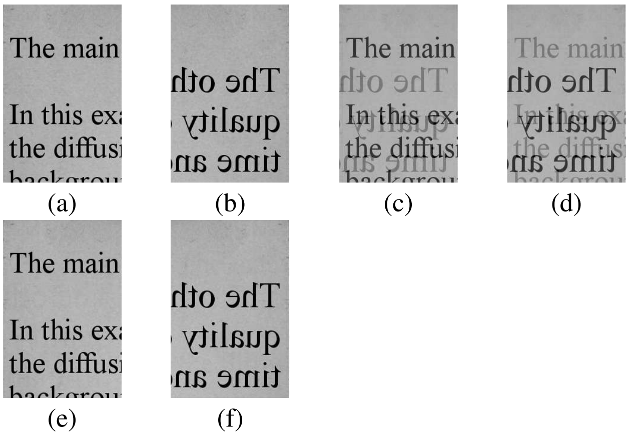


Figure 1: Performance of the ICA method in a linear case. a) and b) source images of the recto and verso sides of the document; c) and d) the degraded images which are the linear combinations of the source images; e) and f) the results of applying the ICA method.

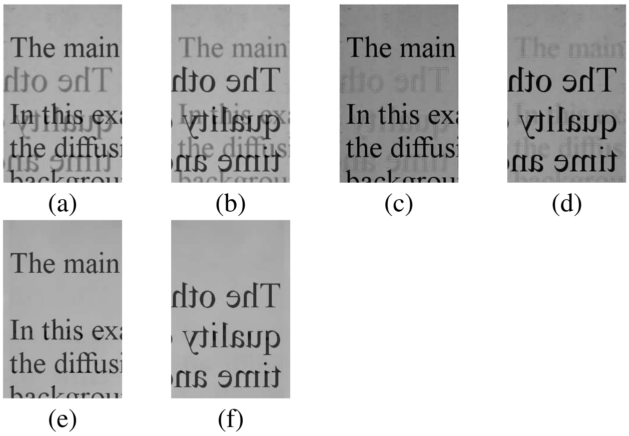


Figure 2: Restoration of bleed-through defect. a) and b) the degraded images generated by the degradation model; c) and d) the results of applying the ICA method; e) and f) the results of applying the diffusion method.

3. PERFORMANCE COMPARISON

There are many implementations for the ICA methods, such as FastICA [14, 15], ICALAB [16], and Symmetric Orthogonalization [4]. The basic ICA method we used is FastICA. As a ground truth test, we applied the code to the samples in [4], and we obtained the same mixing matrix.

In our first example, we used the document images in Figures 1(a) and 1(b) as the defect-free ground truth, and, we used a linear combination to obtain the two degraded images shown in Figures 1(c) and 1(d). In this linear case, the ICA method is able to recover exactly the two sources (see outputs of the ICA method in Figures 1(e) and 1(f)). But similar to the seepage of water or oil in soil, the seepage of ink through a paper is a nonlinear physical phenomenon. Indeed, many parameters such as thickness of document paper and distribution of paper fibers are involved in the seepage phenomenon. There are different models for study of ink seepage in porous media [17, 18]. For example, if we apply the diffusion-based degradation model [18, 19] to the defect-free ground truth in Figures 1(a) and 1(b), we will obtain the two nonlinear degraded images shown Figures 2(a) and 2(b) which show the recto and verso sides of a document. The nonlinear nature of the bleed-through effect can be seen in the figures in the form of horizontal seepage of ink and blur around the interference patterns. The ICA method was applied first, with the results shown in Figures 2(c) and 2(d). For visualization purposes, the outputs are normalized. The sources are well separated, but some weak patterns have remained near the edges of the interference patterns. This is due to one-to-one the nature of the ICA method, which means that it cannot recognize the nonlinear seepage of ink around the patterns. However, all the valuable information is recovered, and the outputs can be passed on to the recognition process following binarization. The outputs of application of our new diffusion method to the input images in Figures 2(a) and 2(b) are shown in Figures 2(e) and 2(f). Because of the nonlinear characteristics of the method, in particular the reverse diffusion processes, there are no marginal patterns. Moreover, the method's background diffusion ensures a uniform background. Another feature of the diffusion methods is that their outputs contain a binarized version of the two sides of the document, which means that the binarization process can be skipped. This is discussed in the next example.

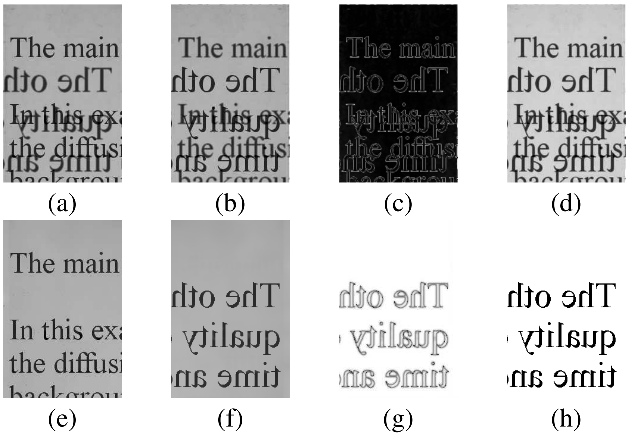

(a) (b) (c) (d)
(e) (f) (g) (h)

Figure 3: Restoration of bleed-through defect of a higher order. a) and b) the degraded images generated by the degradation model; c) and d) the results of applying the ICA method; e) and f) the results of applying the diffusion; g) the corresponding flow field of the verso side of the document; h) the rate of background diffusion to the verso side.

Our third example is more complicated. Here, the rate of seepage is so high that the interference patterns are very close to the main text in terms of gray level. The input images are shown in Figures 3(a) and 3(b), and were obtained by applying the degradation model to the defect-free images in Figures 1(a) and 1(b). The results obtained by the ICA method are shown in Figures 3(c) and 3(d). Because of the fact that the gray levels of the main texts and of the interference patterns are the same, the method separates them from the nonlinear patterns on the edges and boundaries of the strokes. At the same time, because the main text is less dominant than the interference patterns, the diffusion method and, more accurately, its reverse diffusion processes, gradually push the interference patterns to the background. As the interference patterns weaken, they are removed completely by the background diffusion process in the same way as in the first example. However, if the degree of bleed-through becomes so high that the gray levels of the interference patterns are darker than those of the main text, our diffusion method will also fail. This kind of problems is very rare, though, and outside the scope of this work.

Our forth example involves a real sample from the Google Book Search dataset [11] (see Figures 4(a) and 4(b)). The recto side of the document contains not only the main text, but also some handwriting and some weak extra texts. The results of applying the ICA method are shown in Figures 4(c) and 4(d). The interference patterns are very well suppressed and all the overlays have been saved, despite their weak gray levels. This is because of the passive nature of the method, which does not generate any data, but only linearly combines the inputs. However, because of the nonuniform seepage of the ink over the images in this real case, the interference patterns are not completely removed in some regions. The reason for this is that the coefficients in the mixture are assumed to be global and constant, which makes the diffusion method more powerful here. The patterns are removed completely, and the background is very uniform. Although the method can be set up to save very weak and fine structures on the images, its aggressive nature destroys many of the very weak marks.

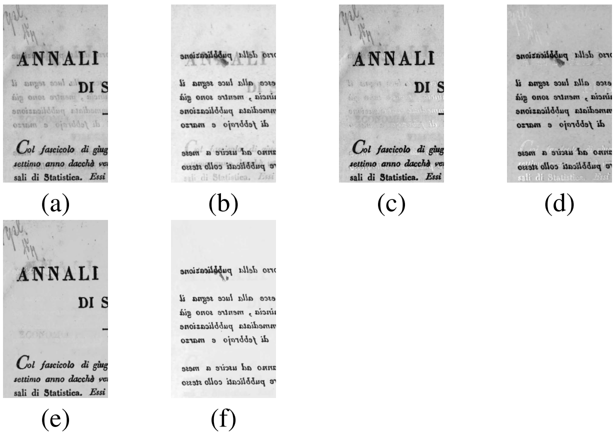

(a) (b) (c) (d)
(e) (f)

Figure 4: Restoration of bleed-through defect. a) and b) the degraded images obtained from the Google Book Search dataset; c) and d) the results of application of the ICA method; e) and f) the results of applying the diffusion method.

All these examples and similar cases reveal several advantages and disadvantages of the two methods. Below, we provide a concise list of these characteristics, which can be used to select the appropriate method in a particular application.

3.1 Advantages of ICA

- **Passive:** The ICA method does not actually add additional information to the input data. The result of ICA processing is, therefore, a near restoration of the true data.
- **Global:** With the ICA method, all the input data are included in the processing and determination of the mixing matrix. Therefore, small and local fluctuations have a little effect on its output.

3.2 Disadvantages of ICA

- **Sensitivity:** There is a one-to-one correspondence in the ICA method between two pixels having the same coordinates on the recto and verso sides of the document. If a spatial shift, due to misalignment in the scanning process, is inserted between the coordinate systems of the images of the two sides, the ICA result will be very poor, and a combination of the images of the recto and verso sides will be extracted, instead of separate images of the two sides. Consequently, the method requires a registration between the images of two sides of the document.
- **Ignorance of 2D spatial correlations:** With the ICA method, the image data is considered to constitute a one-dimensional signal. The two-dimensional relations between gray levels on the image will therefore be reduced to a one-dimensional relation. Consequently, the method does not make use of the information coming from the image nature of the inputs.

3.3 Advantages of the diffusion method

- **Mutual local and global behavior:** The main diffusion processes occur over a small area around each pixel. This local behavior renders the method highly adaptable to local variations. At the same time, the flow field collects information from a more global point of view, causing the method to behave globally.
- **Enhancement: binarization and edges.** The flow field actually represents the edges and continuous boundaries of the strokes. Also, the diffusion rate from the estimated background plate can be used as the binarized version of the document. Moreover, additional types of information can be obtained from other variables, which makes it possible to omit other binarization and edge detection processing tasks.
- **2D correlation:** The data of every pixel is affected by the information of its two-dimensional neighborhood, and so all the correlations between nearby pixels will be used in the process. This helps to save of fine and thin structures in the images.

3.4 Disadvantages of the diffusion method

- **Aggressive:** The most negative aspect is its tendency to alter the values of the image data. In the restoration problems in particular, diffusion-based methods are less applicable.
- **Computational cost:** The computational cost of the diffusion-based method is about ten times higher than ICA. We are working on reducing the computational cost by using other paradigms such as variational framework.

4. NEW DIRECTIONS

In this section, we present some combined methods for the restoration and enhancement of document images based on the results and discussion contained in the previous section.

- **Restoration: Using the diffusion method's results as a guide for ICA.** The coefficients of the source mixture in ICA are global. Here, we modify the coefficients by adding to them a position-dependent variation, resulting in the inclusion of gradual variations in the image and local changes in the source separation. To estimate the image variations, outputs of the diffusion method will be used. In other words, the coefficients of the separation matrix will be a function of position.

  It is interesting to note that a similar idea, which is to apply a diffusion process prior to performing the main task, is used in the wavelet techniques [20].

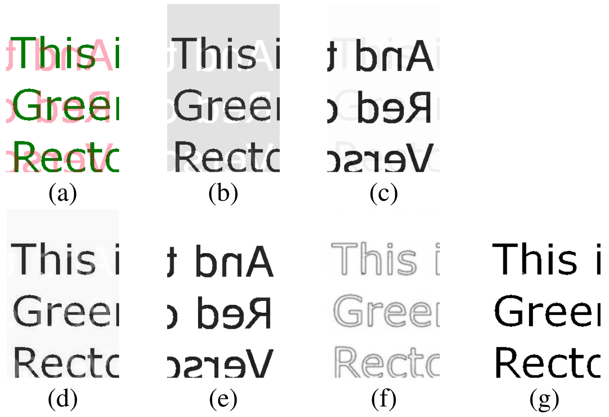
  Another potential approach to including the results of the diffusion method in the ICA method is to redefine the cost function of the ICA method and include a term which computes the distance between the estimated output and the diffusion method results. We are currently working on these approaches.
- **Enhancement: Using ICA results as inputs to the diffusion method.** DFDM is a powerful tool for enhancement and source separation. However, it requires some sort of source dominance in the inputs. Specifically, in each input image one of the sources must be dominant (even at a minor level). In many cases, such as colored single-sided images, the usual RGB channels are not suitable for the diffusion method to be applied. In all three channels, the two sources are very close together in magnitude. In these cases, a pre-separation using ICA will give us two input images which are suitable for DFDM. Then, applying of the diffusion method results in very good enhancement and total separation, even if the first separation task (ICA) fails. Also, the ICA step can be preceded by another diffusion processing which helps in denoising and regularizing the images.

  In Figure 5, for example, a combination of the ICA method and the diffusion method is used to restore a colored single-sided document image. The input image is shown in Figure 5(a). After applying the ICA methods, we have two separate images (see Figures 5(b) and 5(c)). There are some minor marks on the other side on each output. Now, we can apply the diffusion method to the outputs of the ICA method. The results are shown in Figures 5(d) and 5(f). The background is very clear. Also, the flow field of the output presents the continuous boundaries of the text (see Figure 5(g)). Moreover, the diffusion rate from the background plate is equivalent to the binarized version of the text (Figure 5(h)). In other words, there is no need for extra processes to obtain binarized output or edges and boundaries. Although a simple binarization may be performed on Figures 5(b) and 5(c), we recommend DFDM especially to have the results as in Figures 5(f) and 5(g) in the same process. Also, there are other methods for separation in the color space [21, 22], which can be combined with the diffusion method to achieve better results.
- **Content-based information: Using NN to provide content information for the ICA+DFDM combination.** Neural networks (NN) offer the opportunity to classify data based on content of the information. For the pure image processing classifiers, such as the flow field, there is no difference between, say, an accent and a similar mark which does not have any linguistic information. The advantage of NNs is that they can be trained for the recognition or classification of content-based information. In a combination of ICA, NN, and diffusion, the image data will be analyzed at three different levels.

  Although combining DFDM and ICA for the restoration and enhancement of document images is very promising and constitutes a powerful tool, the use of content-related information in the analysis is of great interest, especially in complex cases. In document images that suffer from very variable background and meaningless spots and strokes, this combination will restore or enhance the input document images more easily.

5. CONCLUSION

The advantages and disadvantages of two principal methods (ICA and DFDM) for the restoration of bleed-through defects in very old documents are analyzed. By means of examples, it is shown that the ICA method cannot handle nonlinear phenomena. Also, its highly global nature results in interference patterns remaining when there are gradual variations over the image. However, as a general method, it is a powerful tool. Our new diffusion method (DFDM) behaves in a similar way to the actual physical processes that caused the defect in the first place. It removes the ink seepage appropriately by means of the diffusion method. However, the method requires two input images. Also, it is very aggressive



(a) (b) (c) (d) (e) (f) (g)

Figure 5: Separation of two sides of a document from its colored single-sided document image. a) the input image; b) and c) the results of applying the ICA method to the image in part (a); d) and e) the results of applying the diffusion method to the outputs of the ICA method (images in parts (b) and (c)); f) the corresponding flow field of the recto side of the document, which contains the continuous boundaries if the image; g) the rate of background diffusion to the recto side, which is actually a binarized version of the corresponding text.

and seriously modifies the input data. We therefore propose that, in order to gain all the advantages of these methods, some combination of them be used to restore (or enhance) low quality document images.

REFERENCES

[1] G. Leedham, S. Varma, A. Patankar, and V. Govindaraju, "Separating text and background in degraded document images - a comparison of global thresholding techniques for multi-stage thresholding," in *Proc. 8th IWFHR*, 6–8 Aug. 2002, pp. 244–249.

[2] E. Dubois and A. Pathak, "Reduction of bleed-through in scanned manuscript documents," in *Proc. IS&T Image Processing, Image Quality, Image Capture Systems Conference (PICS2001)*, Montreal, Canada, April 2001, pp. 177–180.

[3] G. Sharma, "Show-through cancellation in scans of duplex printed documents," *Image Processing, IEEE Transactions on*, vol. 10, no. 5, pp. 736–754, 2001.

[4] Anna Tonazzini, Emanuele Salerno, and Luigi Bedini, "Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique," *IJDAR*, vol. 10, no. 1, pp. 17–25, June 2007.

[5] H. Nishida and T. Suzuki, "Correcting show-through effects on document images by multiscale analysis," in *Proc. 16th ICPR*, 2002, vol. 3, pp. 65–68 vol.3.

[6] Hon-Son Don, "A noise attribute thresholding method for document image binarization," *IJDAR*, vol. 4, no. 2, pp. 131–138, Dec. 2001.

[7] Emanuele Salerno, Anna Tonazzini, and Luigi Bedini, "Digital image analysis to enhance underwritten text in the archimedes palimpsest," *IJDAR*, vol. 9, no. 2, pp. 79–87, Apr. 2007.

[8] Xiaowei Zhang, Jianming Lu, and Takashi Yahagi, "Blind separation methods for image show-through problem," in *Information Technology Applications in Biomedicine, 2007. ITAB 2007. 6th International Special Topic Conference on*, Jianming Lu, Ed., Nov. 8–11 2007, pp. 255–258.

[9] Chew Lim Tan, Ruini Cao, Peiyi Shen, Qian Wang, Julia Chee, and Josephine Chang, "Removal of interfering strokes in double-sided document images," in *Applications of Computer Vision, 2000, Fifth IEEE Workshop on.*, Ruini Cao, Ed., 2000, pp. 16–21.

[10] E. Dubois and P. Dano, "Joint compression and restoration of documents with bleed-through," in *Proc. IS&T Archiving 2005*, Washington DC, USA, April 2005, pp. 170–174.

[11] Google, *Book Search Dataset*, Version v edition, 2007.

[12] Gang Zi and D. Doermann, "Document image ground truth generation from electronic text," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, D. Doermann, Ed., 2004, vol. 2, pp. 663–666 Vol.2.

[13] Gang Zi, "Groundtruth generation and document image degradation," Tech. Rep. LAMP-TR-121/CAR-TR-1008/CS-TR-4699/UMIACS-TR-2005-08, University of Maryland, College Park, 2005.

[14] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, June 2000.

[15] E. Oja and Z. Yuan, "The fastica algorithm revisited: Convergence analysis," *Neural Networks, IEEE Transactions on*, vol. 17, no. 6, pp. 1370–1381, 2006.

[16] A. Cichocki, Amari S, Siwek K, T. Tanaka, Anh Huy Phan, and R. Zdunek, "Icalab matlab toolbox ver. 3 for signal processing," 2007.

[17] Xiujin Wang and Jizhou Sun, "The researching about water and ink motion model based on soil-water dynamics in simulating for the chinese painting," in *Image and Graphics, 2007. ICIG 2007. Fourth International Conference on*, Jizhou Sun, Ed., 2007, pp. 880–885.

[18] M. Cheriet and R. Farrahi Moghaddam, "Degradation modeling and enhancement of low quality documents," in *To be appear in WOSPA'2008*, Sharjah, UAE, Invited paper, 2008.

[19] M. Cheriet and R. Farrahi Moghaddam, "Diar: Advances in degradation modelling and processing," in *To be appear in ICIAR'2008*, Póvoa de Varzim, Portugal, Invited paper, 2008.

[20] Yong Yue, M.M. Croitoru, A. Bidani, J.B. Zwischenberger, and Jr. Clark, J.W., "Nonlinear multiscale wavelet diffusion for speckle suppression and edge enhancement in ultrasound images," *Medical Imaging, IEEE Transactions on*, vol. 25, no. 3, pp. 297–311, 2006.

[21] F. Drira, "Towards restoring historic documents degraded over time," in *Document Image Analysis for Libraries, 2006. DIAL '06. Second International Conference on*, 2006, pp. 8 pp.–.

[22] Yingzi Du, Chein-I Chang, and Paul David Thouin, "Unsupervised approach to color video thresholding," *Opt. Eng.*, vol. 43, no. 2, pp. 282–289, Feb. 2004.